オントロジーを用いた文書の意味空間へのマップによる知識型計量分析手法

代表研究者　　山　田　智　子　　宮城大学大学院　事業構想学研究科　博士後期課程
共同研究者　　富　樫　　敦　　宮城大学　事業構想学部デザイン情報学科　教授

1 序論

1-1 研究目的

本研究では，近年，知識処理システム構築の分野で盛んに研究が行われている「オントロジー技術」を用い，語彙の体系化を行い，オントロジーとして構築された概念空間場に文書をマップする（投影する）ことにより，文書に潜む概念構造や意味構造を自動的に獲得するための方法論の確立を目指す.

本研究では，近年，知識処理システム構築の分野で盛んに研究が行われている「オントロジー技術」を用い，語彙の体系化を行い，オントロジーとして構築された概念空間場に文書をマップする（投影する）ことにより，文書に潜む概念構造や意味構造を自動的に獲得するための方法論の確立を目指す.

従来，文書の解析ではテキストマイニング技術が使われるが，テキストマイニングでは形態素の表層的情報を中心とした解析を用いており，語彙や語彙間の意味的関係を把握できないという課題がある．しかし，意味的関係を考慮することで，文書解析の質の向上が可能になると考えられる.

そこで本研究では，文書の内容を把握するためには，文書中の個々の語彙表現の意味とそれらの相互関係を分析することが必要であると考え，対象ドメインの語彙をオントロジーとして体系化することで，分野を特徴付ける語彙（キーワード）を空間的に配置し，文書の意味構造を解析するための基盤となる汎用の概念空間を構築する．文書の解析にオントロジーを用いることで，解析対象となる文書が単なる語彙の集まりとしてではなく，文書全体で大きな意味を持ったデータとして扱われ，各文書について統一的な付加情報をもたせることが可能になり，文書解析に本当に必要な情報を的確に抽出することが可能となる.

その結果，従来ゼロ次元的構造しか持ち得なかったキーワード群が，本研究の成果により，概念間の階層的相対関係，類似性，背反関係等のメタな概念情報が付加された，汎用かつ応用性の高い豊潤な高次元的意味空間として生まれ変わる．本研究では，文書に出現する語彙群を類似性を考慮して抽出し，これらからなる汎用の概念空間の部分空間をその文書の意味として定式化する．この手法により，個々の文書間の関係が基準となる概念空間上での関係として表現でき，それらの相対関係を定性的かつ定量的に計量することで，文書の類似性やその度合い（類似度）さらには包含性や中心性といった文書のネットワーク分析で必須となる主要概念を従来手法と比較してより高次にかつ的確に定式化することが可能となる.

その成果を使い，現在，取り組んでいる研究テーマ「サービス科学における循環型サービス設計モデルの形式化・概念化・理論化」に貢献すべく，本研究を推進していく.

1-2 研究実施内容

本研究では，オントロジーを文書解析に用いた新しい解析手法を提案し，従来手法と比較することで本手法の有効性・妥当性を実証する．本研究では，分析対象分野を科学技術の中でも「情報通信」を対象に行った.

また，本研究で提案する分析手法を文書の解析以外での適用も検討する，具体的には近年，研究が盛んに行われている「サービス科学（サービス・サイエンス，Service Science，Service Science and Engineering)」分野での適用可能性も検討する.

２　研究概要

従来，文書の解析にはテキストマイニングを用いることが多いが，テキストマイニングでは形態素の表層的情報を中心とした解析を用いており，語彙や語彙間の意味的関係を把握できないという課題がある．現在，自然言語処理やデータマイニングの分野で「意味解析」を用いた進化したテキストマイニングが研究されているが，まだ有用な手法が確立されていない現状がある．

そこで本研究では，文書の解析に意味的関係をもたせることに着目した．意味的関係をもたせるためにオントロジーを用いて文書の解析を行うことで意味的関係を把握することが可能なると考えた．

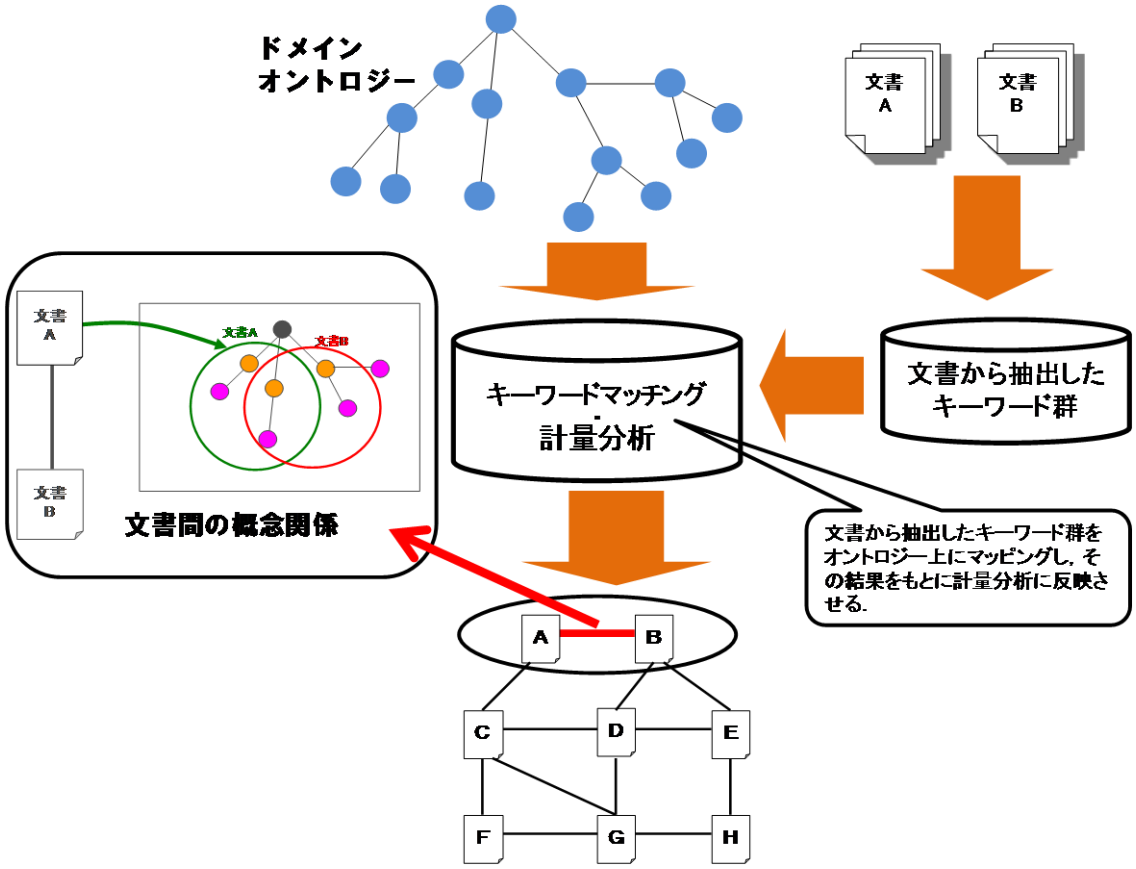

図 1　研究概要図

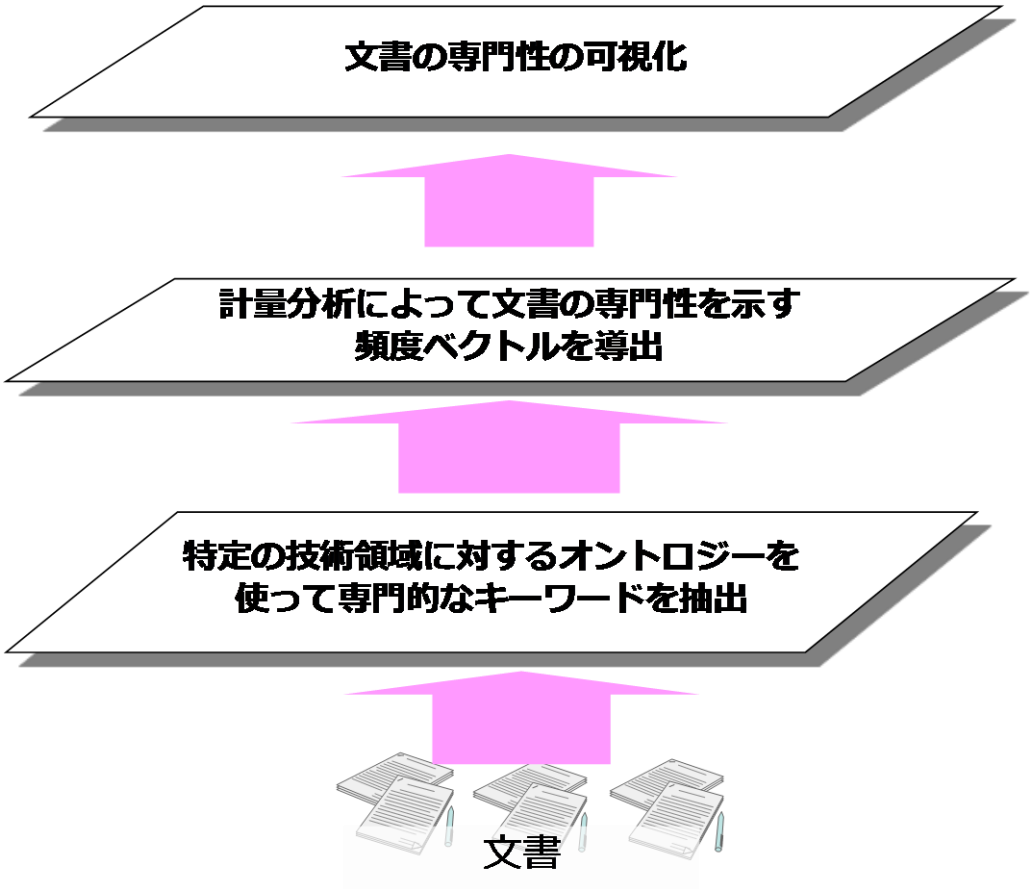

図 2　分析のプロセス

3　分析対象分野語彙抽出

本研究で提案実証する分析手法では，各科学技術分野の特徴を語彙（当該分野を代表する専門用語）で表し，対象文書の分析で用いている．本研究では，分析の精度を質的に高めるため，語彙の作成に情報工学や人工知能分野の新技術「オントロジー技術」の概念を取り入れ，科学技術各分野の語彙の体系化を行い，分析の信頼性を高める．

3-1　専門用語を適切に抽出することの重要性

テキストマイニングでは，分析対象文書に対して，情報工学の自然言語処理分野で使われている形態素解析ツール「茶筅」を使い，文書のデータ解析を行い，その結果に基づいてキーワード群の抽出を行う．文書に出現するキーワードの出現頻出を調べることは，計量分析において，文書の構造化・可視化の前提となる，文書解析を行う上で必須となる作業である．同論文では，web 上に公開されているテキストマイニングフリーソフトウェア「KH Coder」[26]を利用し，

しかし，キーワード抽出のステップでは，形態素解析ツール「茶筅」を使用してキーワードの抽出では，形態素解析の限界（表層的情報）があり，各分野の特徴付けるキーワードが抽出できず，分析の精度が低下することがわかった．

本研究では，茶筅を使い，形態素解析をし，その結果から計量分析・可視化を行っていたが，一部，結果を検証したところ，うまく説明できない関連がなされていることがわかり，適切にキーワードが抽出できていないとの結論に至った．本研究の分析では，専門用語辞書の品質が分析結果に大きく依存するが，専門用語辞書を構築しようとすると，さまざまな専門家の時間的・金銭的なコストがかかり難しいため，手動で非専門家が分野依存の辞書をつくることは不可能に近いのではないかと考えられる．

3-2　専門用語を適切に抽出する際の課題

テキストマイニングでは，キーワード抽出に関して，形態素解析ツール「茶筅」を使い，キーワード群を作成する．

しかし，茶筅だけでは適切な語彙の抽出が難しいことがわかった．例えば，「情報通信技術」という語彙を抽出したい場合，茶筅のみだと「情報，通信，技術」と３つの語彙に分割され，適切な語彙にならないという事がわかり，茶筅＋α という方法が必要であるとの結論に達した．

3-3　専門用語の理想的な抽出法

本研究では，専門用語辞書の品質が分析結果に大きく依存するため，適切な専門用語辞書の構築（語彙の体系化）が必要であるが，専門家に依頼をすると構築に多大なコストがかかる．

そこで，専門家の代行として学会が制作したハンドブックを利用して，専門用語辞書の構築（語彙の体系化）に貢献できるのではないか，と予測した．学会が制作したハンドブックは制作に携わっている方々が学会会員（学会の対象領域の専門家）であることから，専門用語の構築（語彙の体系化）の信頼性を高める方法であると検討した．

3-4　情報通信技術の概念体系を示すオントロジーの構築

本研究では，電子情報通信学会が制作・出版している「電子情報通信ハンドブック」を利用し，語彙の体系化を行い，分析に利用した．

具体的には，電子情報通信学会ハンドブック委員会 編「電子情報通信ハンドブック」の構成に基づいて情報通信に関連する技術を９つの「分野（群）」に分け，各分野はさらに複数の「編」で構成される「編」の中で説明されている技術キーワードとそれらの類義語，同義語を，11,482 語の単語からなる概念体系を構築した．

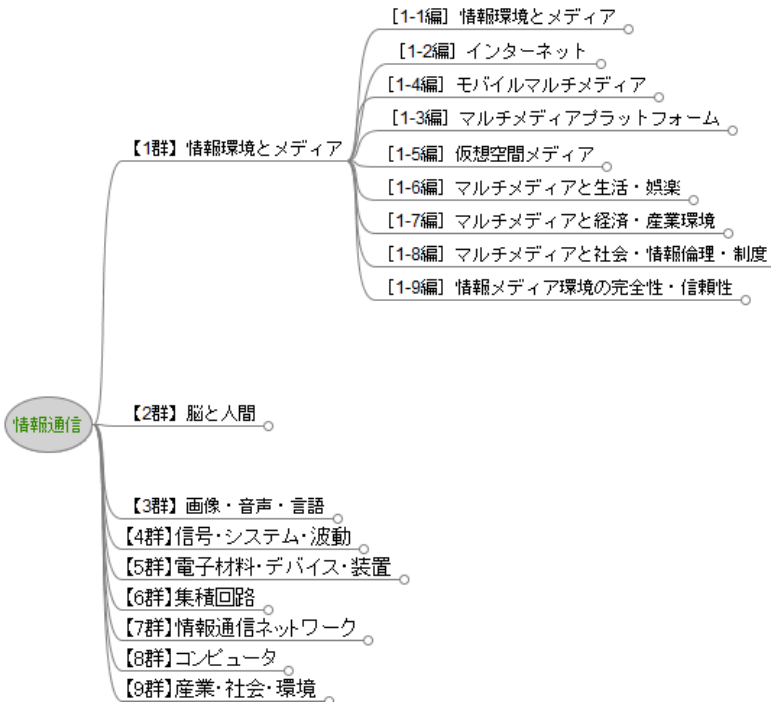

図 3　電子情報通信ハンドブックを使った語彙の体系化の考え方

４　オントロジーを用いた文書の特徴分析

本研究では，文書解析の質の向上を図るために文書中の個々の語彙表現の意味とそれらの相互関係を分析することが必要であると考えた．そこで，オントロジーを使い語彙の体系化を構築することとした．オントロジーとは，タクソノミーに概念関係が追加されたものであり，本研究での，文書の意味構造を解析するための基盤となる汎用の概念空間を構築が可能になると考える．

本研究では，「情報通信」分野を対象に特徴分析を検討した．具体的には，文書から抽出したキーワードに対して，電子情報通信ハンドブックを利用して構築した情報通信分野の語彙の体系上にマッチングを行う．オントロジーでは，上位層が一般的な概念で下位層は概念の具体化という考えから，本研究では，体系化した階層毎に重みを付加する．下位層にいくほど，概念が具体的になることから，本研究の分析では，下位層にいくほど分野に特化した語彙になるという考えと捉え，語彙の重みが大きくなるように設定する．オントロジー上に投影されることで，文書の概念構造や意味構造の獲得が可能になる．

オントロジーに投影した結果をもとに，類似性やベクトル空間法などをベースに計量分析を行い，文書間の関連性を計算する．

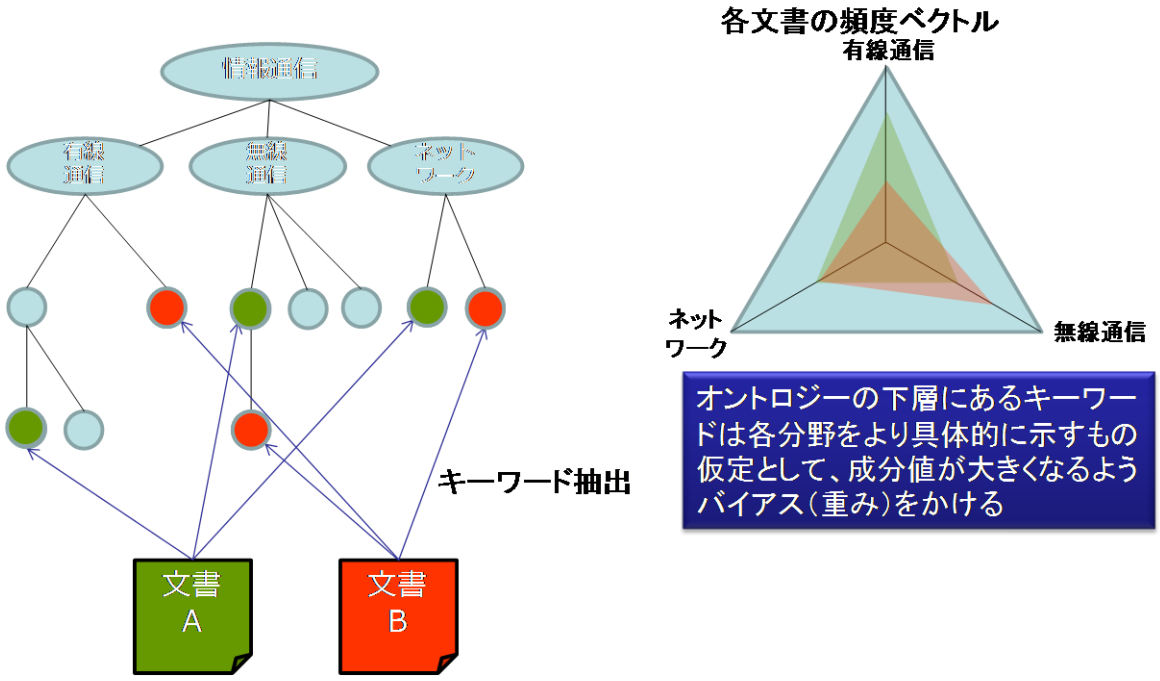

図 4　オントロジーを用いた文書の特徴分析概念図

５　サービス科学の視点からの研究成果の実践的応用

本研究では，得られた成果をビジネス的実践的応用に取り組んでおり，現在，取り組んでいる研究テーマ「サービス科学における循環型サービス設計モデルの形式化・概念化・理論化」に貢献すべく，本研究を推進していく．

サービス科学（サービス・サイエンス，Service Science，Service Science and Engineering）とは，経営工学，社会工学，システム科学，生産管理，マーケティング・サイエンス，法律学，経営戦略などをはじめとする様々な学術分野が融合し，サービスについての研究を行う新しい領域の学問である．

学問の垣根を越えてサービス・システムを推進し，人々の知恵と技術を駆使し，他者へ価値を提供する複雑なシステムを仕様化することである．より正確に述べるなら，サービス科学とは，科学，経営，工学の応用的領域であり，企業など組織が他者と一体となり利益（価値）を生み出すことである．

サービス科学の目的として，サービスに対して，科学的な手法，効率的なマネジメント手法，生産性を最大限に高めるための工学的な生産方法を提供し，サービスの特性（同時性・消滅性，不可分性，不均質性，非有形性）に起因するところの諸問題を解決し，生産効率をあげ，また，イノベーションをシステマチィックに生み出す枠組みを見出すことが課題になっている．

本研究では，得られた成果を使い，「サービス科学における循環型サービス設計モデルの構築」を検討していく．

また，今日，先進国だけでなく，急速に発展が進む開発途上国においても，サービス産業は経済的に最も大きい割合を占めており，これらサービス産業が、よりシステム工学的なイノベーションを行うことに，学術界，産業界，政府らは注目している．サービス科学は，今までの学術領域の再編というわけではなく，多くの学術領域をまたがって，サービスに関する知恵を集め，新たな学術領域，研究分野として確立されることが望まれている．

サービスは，①同時性・消滅性，②不可分性，③不均質性，④非有形性，という特性があり，この特性の解決方法として，情報技術や知識処理などの情報科学を用いて研究がされている．

６ サービス科学に対する背景と問題設定

6-1 サービス科学に対する背景

我が国の産業構造は，他の先進諸国同様サービス産業に大きく依拠している（2009 年度経済産業省通商白書のデータとして，全産業に対するサービス産業の占める割合は約 72%である）割には，サービス産業の生産性の伸びは製造業に比較して相対的に小さい（同通商白書によると，サービス産業の生産性は全産業に対して GDP 比で約 47%である）．2004 年の IBM によるサービスサイエンス（その後，サービスサイエンス・マネージメントエンジニアリング：SSME と改称）の提唱，IBM の提唱よりさらに過去に遡る吉川氏らのサービス工学の提唱により，最近では経営分野や知識工学分野等でサービスに関し盛んに議論されてはいるが，コンピュータサイエンスと比較すると，未だに経験と勘に頼る部分が多く，学問としての体系化を形成していない．学問分野としての「サービス科学」を確立するためには，再現性，普遍性，客観性などの特徴を有するサービスの形式的モデル化，理論化を行うことが喫緊の最重要課題である．

サービスの理論的モデル化が進まない原因に，サービスの対象をモノ作り，マーケティング，イノベーションに終始し，サービスを形式的に余り議論していない点にある．つまり，サービスを質と量の側面から十分かつ形式的に議論していないことが最大の原因である．多くの議論は経営的，人文社会学的であり，自然科学的な形式的議論が欠けている．

以上の問題を克服しサービスの継続的改善と設計を目指したモデルに，サービス工学でのサービスの最適設計モデルがある．サービスの提供者，受容者，サービスの改善を意識した観測者，設計者を明示し，サービスのフローと設計に関する枠組みを与えた極めて有用な概念モデルである．この循環モデルは，信頼性工学のデミングサイクル PDCA（Plan Do Check Action）サイクル，及び製造分野での PLM（Product Lifecycle Management）と密接な関係があり，サービスを科学的に体系化する上での布石と成り得る．

しかし，この循環モデルにも問題が山積する．そのモデルの詳細化，形式化に関しては今後の十分な議論と研究を待たねばならない．概念に終始しないサービス（価値）提供者，受容者の形式モデル（研究クラスターのヒューマンモデリングに相当），サービスの質的・量的基準を含むサービスの形式的詳細モデル（同プロセスマネジメントに相当），サービス提供者と受容者とのやりとりの形式化と観測結果の分析手法（同サービス値に相当），分析結果からの構成的サービスの改善手法，再構築されたサービスのシステマチックな提供手法（同サービス創生社会のモデリングに相当）が形式的に与えられなければならない．

本研究では，このサービス設計の巡回モデルに科学的展開としての大きな意義を見出し，モデルに内在する問題の解決，理論的詳細化，実証実験を通した理論化の実践的検証を通して，その形式的モデル化の妥当性を検証し試みる研究を行っている．

6-2　当該研究におけるサービス科学の問題設定

サービス研究に関し，普遍性・再現性・客観性を有するサービスの汎用的形式モデルと理論が構築できていないことが大きな課題である．これまで，サービスの汎用モデルが確立していなかったために，分野に応

じて未だに経験と勘に頼る百鬼夜行的議論が横行し，サービスの本質を突いた共通な議論ができず仕舞いになっていた．汎用モデルが構築できることにより，枝葉末節には左右されないサービスがカバーする全分野を俯瞰する議論の展開が可能となる．以上は，余りに大きい問題であり，以下に少しブレークダウンした問題設定を列挙する．本研究では，「サービスの設計に限定した汎用的形式モデルと理論を構築すること」を課題として設定する．

（問題 1）サービス設計に関する包括的循環モデルの詳細が十分検討されていないこと．（◎）
（問題 2）サービス提供者，受容者に対する形式モデルが確立していないこと．（○）
（問題 3）質的・量的基準を含むサービスの形式的詳細化モデルが確立していないこと．（△）
（問題 4）サービス提供者と受容者とのやりとりの形式化と観測結果の分析手法が不十分なこと．
（問題 5）分析結果からの構成的サービスの改善手法，再構築されたサービスのシステマチックな提供手法が形式的に議論されていないこと．

本研究では，問題１を重点課題とし，一部問題２，問題３を関連課題として取り組む．
（本研究での，◎：重点的解決問題，○：解決問題，△：関連問題，無印：参考問題．）

6-3　問題設定のまとめ

サービス研究の問題設定に関し，次のようにまとめることができる．

(a) (6-2)で設定した問題が解決された場合の具体体なイメージとしては，汎用モデル・理論により，サービスを鳥瞰的（マクロ的）に見渡せる理論的なフレームワークが構築できることが挙げられる．電磁気におけるマックスウェルの方程式，相対性理論の効用・意義と等しく，サービスを大局的に把握することができる．個別ドメインのサービスを扱う場合は，モデル・理論を特化し，ドメイン特有の例外にフォーカスできる利点がある．（設定問題が解決された場合のイメージ）

(b) サービスの理論化は，散在するサービス論議を統一，一元化する効果があり，その効果はすべてのサービスに及ぶ．

7　サービス科学の視点からの文書解析とオントロジーの役割

サービス科学で検討しているモデルの詳細化，形式化に関しては今後の十分な議論と研究を待たねばならない．サービスに終始しないサービス（価値）提供者，受容者の形式モデル，サービスの質的・量的基準を含むサービスの形式的詳細モデル，サービス提供者と受容者とのやりとりの形式化と観測結果の分析手法，分析結果からの構成的サービスの改善手法などが形式的に与えられなければならない．

そこで本研究で得られた成果，文書の解析に知識的要素をサービス提供者と受容者の観測・分析に応用することで，サービス提供者・受容者それぞれの指向や動向を把握することが可能になり，目的にあったサービスを提供することが可能になると考える．

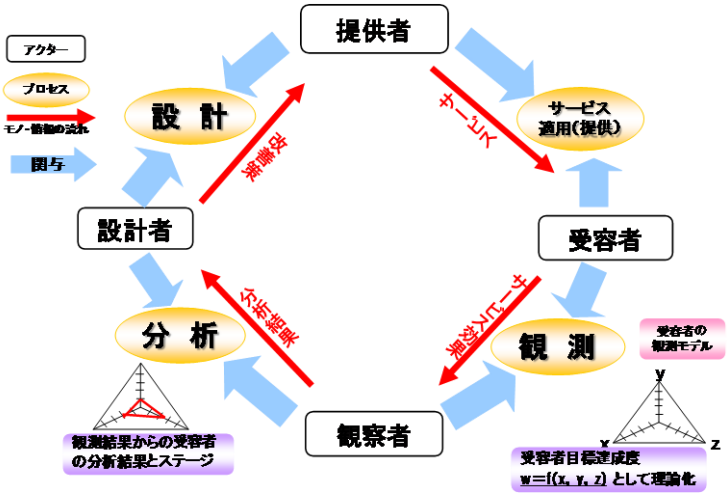

図 5　現在，取り組んでいる「サービス科学における循環型サービス設計モデル」概念図

８　まとめ

本研究では，近年，知識処理システム構築の分野で盛んに研究が行われている「オントロジー技術」を用い，語彙の体系化を行い，オントロジーとして構築された概念空間場に文書をマップする（投影する）ことにより，文書に潜む概念構造や意味構造を自動的に獲得するための方法論の確立を目指した．

本研究の分析では，語彙群の品質が分析結果に大きく依存するため，語彙の抽出に関して，領域オントロジーの概念を取り入れ，適切な語彙の抽出を行った．

また，分析のために情報通信分野に関する語彙の体系化を行い，文書の特徴分析に適用可能であることがわかった．

そして，本研究で得られた分析手法に関して，実践的応用としてサービス科学での視点において，現在，検討中のサービス循環モデルに対して，観測・分析部分において応用可能性の期待できることがわかった．

【参考文献】

[1]山田智子, 藤井章博「科学技術政策に関する研究」研究・技術計画学会　第 21 回年次学術大会 2006.10

[2]山田智子, 富樫敦, 藤井章博「科学技術政策を対象とした政策調査研究資料の計量的分析」情報処理学会第 69 回全国大会 2007.3

[3]山田智子, 坂本眞一郎, 藤井章博「科学技術予測調査文書データベースの視覚的分析」日本情報経営学会第 55 回大会 2007.11

[4]山田智子, 富樫敦, 藤井章博「科学技術予測調査文書の視覚的分析」情報処理学会第 70 回全国大会 2008.3

[5] 西城英之, 鈴木智充, 山田智子, 富樫敦「ベクトル空間法を基礎としたカテゴリーマッピング法による Web ページの自動分類」情報処理学会研究報告, マルチメディアと分散処理研究会報告 (2008 年 06 月)

[6]山田智子：日本科学協会笹川科学助成・研究完了報告書(2008 年度)“オントロジーを用いた知的文献計量分析手法の研究”, 2009 年 3 月

[7]山田智子, 富樫敦 他: (財)新技術振興渡辺記念会･科学技術調査研究助成課題報告書(2008 年度)“科学技術政策を対象とした文書の知識型計量分析手法の研究”, 2009 年 3 月.

[8] 溝口祐美子, 中本利明, 浅川一満 他「オントロジーを用いた文書間類似度計算手法」電子情報通信学会技術研究報告. AI, 人工知能と知識処理 108(119), 87-92, 2008-06-23

[9] 溝口祐美子, 長野伸一, 稲葉真純 他「オントロジーを利用した文書間のセマンティックな類似度計算手法」電子情報通信学会技術研究報告. AI, 人工知能と知識処理 109(51), 1-6, 2009-05-15

[10]Takeshi Morita, Noriaki Izumi, Naoki Fukuta, Takahira Yamaguchi, ″A Graphical RDF-based Meta-Model Management Tool″, IEICE Transactions on Information and Systems, Vol.E89-D No.4 pp.1368--1377, (2006.4)

[11] Takeshi Morita, Naoki Fukuta, Noriaki Izumi, Takahira Yamaguchi, ″DODDLE-OWL: A Domain Ontology Construction Tool with OWL″, Proceedings of the 1st Asian Semantic Web Conference Lecture Notes in Computer Science Vol.4185 pp.537--551, (2006.9)

[12] RADA R. Development and application of a metric on semantic nets. IEEE Trans. System, Man, and Cybernetics 19, 17-30, 1989

[13] LI Y. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering 15(4), 871-882, 2003

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
文書の知識型計量分析手法の提案と企業情報システムへの適用	日本ビジネス・マネジメント学会第 9 回全国研究発表大会	2012 年 8 月